Probability theory is nothing but common sense reduced to calculation. — Pierre Laplace, 1812

Reasoning under Uncertainty.

Probability

Luis G. Jaimes

How do we quantify uncertainty?

Uncertainty is quantified using **Probability**.
Instead of saying:
"This might happen", we assign a number:

→ event A happens with 80% chance

Probability distribution (intuition)

A probability distribution assigns likelihoods to all possible outcomes:

Example (coin):
Heads: 0.5
Tails: 0.5
This fully describes the uncertainty

Example (rolling a dice)
(1,2,3,4,5,6)
Each outcome has:

How do we measure a probability distribution?

A probability distribution itself is not just “uncertain”; it has a **structure**.

We measure it using:

Spread

How concentrated or spread out the probabilities are

Variance

Measures how far values deviate from the mean

Entropy (how much uncertainty is in a distribution**)**

Entropy

Entropy is a concept from **Information Theory**.
It measures:

how much uncertainty is in a distribution

Example

Let’s compute the entropy for the coin:

Compute the H(X) = ?

Example

Let’s compute the entropy for the coin:

Compute the $H=-[0.5\log2(0.5)+0.5\log2(0.5)] = H=-[-0.5-0.5]$,

This is the maximum entropy for a binary variable
The coin is completely uncertain
You gain 1 bit of information when you observe the result

Example 2

Given the following distribution (0.2,0.5,0.3), measure its entropy:

Example 2

Given the following distribution, measure its entropy: (0.2,0.5,0.3)

$H = -[0.2\log_2(0.2) + 0.5\log_2(0.5) + 0.3\log_2(0.3)] = \approx 1.49$ bits

Good or bad ?

Maximum entropy p1=p2=p3= 1/3, (1/3, 1/3, 1/3)
H(X) = 1.585,

So, 1.49 < 1.585, You can “guess better” than random

Example 3

Compute the entropy of the following, probability distribution:

What is the meaning of that?

Lower, and upper bounds
$0 \leq H(X) \leq \log_2(n)$

Discrete random variables

A variable that represents the outcome of a random process

- Example: Random variable X could be weather, taking any of the following possible values:

{**sunny**, **rainy**, **snowy**, **cloudy**, **partly cloudy**, and **windy**}

- In this example, these are **distinct categories**, not continuous numbers

Notation

What is ?

This means:
- “What is the probability that X takes the value x?”
{**sunny**, **rainy**, **snowy**, **cloudy**, **partly cloudy**, and **windy**}

Example

p(x) or p(sunny) = 4

A Probability Mass Function (PMF) countable outcomes

A function that assigns a probability to **each possible value** of a discrete random variable:

The possible outcomes are countable, like {**sunny**, **rainy**, **snowy**, **cloudy**, **partly cloudy**, and **wind**

Example:

Weather	Probability
sunny	0.4
rainy	0.2
cloudy	0.2
windy	0.1
snowy	0.05
partly cloudy	0.05

Probability Density Function (PDF). Here, instead of weather, the variable could be temperature.

Example: Temperature (continuous version of weather)

The value is **not a probability**
Instead:

Important fact

Probability Density Function PDF

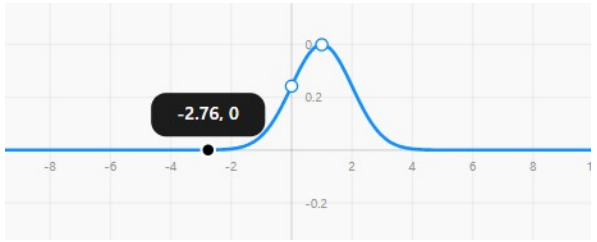
A classic example is the **Normal Distribution**.

X=temperature in °C Values like: 23.1, 23.11, 23.111, ... Infinite possibilities

The value is **not a probability,**

Instead:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

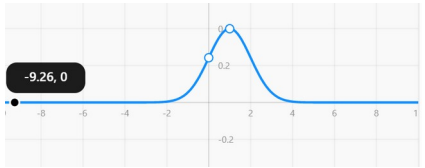


Probability Density Function PDF

What is f(5)= ?

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The f (5) is the height of the curve at



If you want the probability of 5, , then you can do P(4.9≤X≤5.1)

Probability = “how much area is under the curve near 5”
Plugging gives you a **density value**, not a probability. How “concentrated” probability is around 5

Two views of probability

- **Frequentist:** Probability = long-run frequency of events

$$P(A) = \lim_{n\to\infty} \frac{\#\text{ times A happens}}{n} \qquad P(\text{heads}) = \lim_{n\to\infty} \frac{\#\text{ heads}}{n}$$

- **Bayesian** interpretation of probability. Probability = **degree of belief (uncertainty)**

Example: P(ice cap melts)=0.7, which means, based on current knowledge, we are 70% confident

Bayesian probability changes with evidence:

Discrete random variables and notation

- Example: "A" might be the logical expression "it will rain tomorrow".
- $p(A)$ denotes the probability that the event A is true.

- Some properties
  - $0 \leq p(A) \leq 1$
  - $p(A) = 0$ means the event definitely will not happen
  - $p(A) = 1$ means the event definitely will happen.
  - $p(\sim A) = 1 - p(A)$.

Discrete random variables and notation

- This is the world of pmf

- PROBABILITY MASS FUNCTION

Discrete random variables, probability mass function pmf

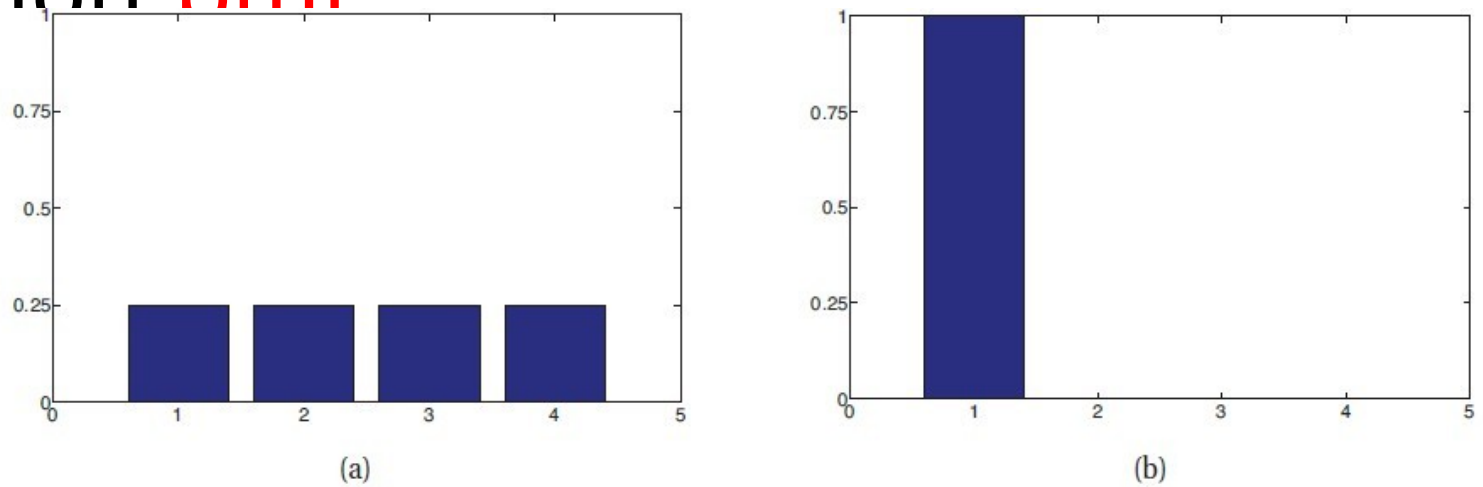


Figure 2.1 (A) a uniform distribution on $\{1, 2, 3, 4\}$, with $p(x = k) = 1/4$. (b) a degenerate distribution $p(x) = 1$ if $x = 1$ and $p(x) = 0$ if $x \in \{2, 3, 4\}$. Figure generated by `discreteProbDistFig`.

This satisfies the properties $0 \leq p(x) \leq 1$ and $\sum_{x \in \mathcal{X}} p(x) = 1$.

- **Probability of a union of two events**

$p(A \vee B) = p(A) + p(B) - p(A \wedge B)$ OR $p(A) + p(B)$ if A and B are mutually exclusive

- **Joint probabilities (INTERESTING STUFF)**

$p(A,B) = p(A \wedge B) = p(\mathrm{A}|\mathrm{B})p(B)$

- **Marginal distribution:** Given a **joint distribution** on two events $p(A,B)$, we define the **marginal distribution** as follows:

Marginal distribution

Given a joint distribution over two variables and , written as:

The marginal distribution of one variable is obtained by summing or integrating out the other variable.

Discrete Case
If and are discrete:

You sum over all possible values of the other variable.

- **chain rule** of probability or join probabilities

$$p(X_{1:D}) = p(X_1)p(X_2|X_1)p(X_3|X_2, X_1)p(X_4|X_1, X_2, X_3)\dots p(X_D|X_{1:D-1})$$

where we introduce the Matlab-like notation $1 : D$ to denote the set $\{1, 2, \dots, D\}$.

- of event *A*, given that event *B* is true, as follows:
- **Conditional probability**

$$p(A|B) = \frac{p(A, B)}{p(B)} \text{ if } p(B) > 0$$

- **Bayes rule**

$$p(X = x|Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)} = \frac{p(X = x)p(Y = y|X = x)}{\sum_{x'} p(X = x')p(Y = y|X = x')}$$

Examples (Join probability, join distributions)

- Example: lets say we want to estimate students' success by observing three DISCRETE random Variables: Intelligence (I), difficulty of the classes the student takes (D), and Grades(G) .
- P(I,D,G) join probability distribution.

- Intelligence (I) ← 2
  - i^0 (low), i^1 (high),
- Difficulty (D) ← 2
  - d^0 (easy), d^1 (hard)
- Grade (G) ← 3
  - g^1 (A), g^2 (B), g^3 (C)

parameters
$2 \times 2 \times 3 = 12$

independent params
11

I	D	G	Prob.
i^0	d^0	g^1	0.126
i^0	d^0	g^2	0.168
i^0	d^0	g^3	0.126
i^0	d^1	g^1	0.009
i^0	d^1	g^2	0.045
i^0	d^1	g^3	0.126
i^1	d^0	g^1	0.252
i^1	d^0	g^2	0.0224
i^1	d^0	g^3	0.0056
i^1	d^1	g^1	0.06
i^1	d^1	g^2	0.036
i^1	d^1	g^3	0.024

1

conditional probability

- P(I,D,G | $G=g^1$)
  - Question for 5 points. Is that Table a distribution?

I	D	G	Prob.
i^0	d^0	g^1	0.126
i^0	d^0	g^2	0.168
i^0	d^0	g^3	0.126
i^0	d^1	g^1	0.009
i^0	d^1	g^2	0.045
i^0	d^1	g^3	0.126
i^1	d^0	g^1	0.252
i^1	d^0	g^2	0.0224
i^1	d^0	g^3	0.0056
i^1	d^1	g^1	0.06
i^1	d^1	g^2	0.036
i^1	d^1	g^3	0.024

CONDITIONING, P(I,D,G | G=g^1)

- Question for 5 points.

  Is that Table a distribution?

I	D	G	Prob.
i^0	d^0	g^1	0.126
i^0	d^0	g^2	0.168
i^0	d^0	g^3	0.126
i^0	d^1	g^1	0.009
i^0	d^1	g^2	0.045
i^0	d^1	g^3	0.126
i^1	d^0	g^1	0.252
i^1	d^0	g^2	0.0224
i^1	d^0	g^3	0.0056
i^1	d^1	g^1	0.06
i^1	d^1	g^2	0.036
i^1	d^1	g^3	0.024

Why not?

Because

$$\sum_{x \in \mathcal{X}} p(x) \neq 1.$$

I	D	G	Prob.
i^0	d^0	g^1	0.126
i^0	d^0	g^2	0.168
i^0	d^0	g^3	0.126
i^0	d^1	g^1	0.009
i^0	d^1	g^2	0.045
i^0	d^1	g^3	0.126
i^1	d^0	g^1	0.252
i^1	d^0	g^2	0.0224
i^1	d^0	g^3	0.0056
i^1	d^1	g^1	0.06
i^1	d^1	g^2	0.036
i^1	d^1	g^3	0.024

I	D	G	Prob.
i^0	d^0	g^1	0.126
i^0	d^1	g^1	0.009
i^1	d^0	g^1	0.252
i^1	d^1	g^1	0.06

Normalized (divide by the summation)

I	D	G	Prob.
i^0	d^0	g^1	0.126 / 0.447
i^0	d^1	g^1	0.009
i^1	d^0	g^1	0.252
i^1	d^1	g^1	0.06

$P(I, D, g^1)$

unnormalized meas

0.447

I	D	Prob.
i^0	d^0	0.282
i^0	d^1	0.02
i^1	d^0	0.564
i^1	d^1	0.134

$P(I, D \mid g^1)$

Example: Marginalization

$$p(A) = \sum_b p(A, B) = \sum_b p(A|B = b)p(B = b)$$

- *Given the join distribution* $p(I, D)$, *Marginalize I, namely* $p(D) = \sum_I p(I, D) = \sum_I p(D|I)p(I)$

$p(I) = \sigma_D\, p(I, D) = p(I, D = d^0) + p(I, D = d^1)$

I	D	Prob.
i^0	d^0	0.282
i^0	d^1	0.02
i^1	d^0	0.564
i^1	d^1	0.134

D	Prob.
d^0	0.846
d^1	0.154

Example: Expand the summation

$\sum_I P(I, D) = P(D)$. As a concrete example of marginalization, say you have throw two 6-sided dice, D_1 and D_2. This defines a joint probability distribution $P(D_1, D_2)$. The probability that $D_2 = 1$ is equals to

$\sum_{i=1}^{6} P(D_1 = i, D_2 = 1)$, since D_1 can only take on the values 1 to 6.

The joint distribution

P(D1, D2)=

D_1	D_2	$P(D_1, D_2)$
1	1	1/36
1	2	1/36
1	3	1/36
...	...	...
6	6	1/36

$$\sum_{i=1}^{6} P(D_1 = i, D_2 = 1)$$

$$P(1,1) + P(2,1) + P(3,1) + P(4,1) + P(5,1) + P(6,1)$$

$$P(D_1 = 1, D_2 = 1) + P(D_1 = 2, D_2 = 1) + P(D_1 = 3, D_2 = 1)$$
$$+ P(D_1 = 4, D_2 = 1) + P(D_1 = 5, D_2 = 1) + P(D_1 = 6, D_2 = 1)$$

$$\frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} \quad = \frac{6}{36} = \frac{1}{6}$$

Independence and conditional independence

Independence and conditional independence

We say X and Y are **unconditionally independent** or **marginally independent**, denoted $X \perp Y$, if we can represent the joint as the product of the two marginals (see Figure 2.2), i.e.,

$$X \perp Y \iff p(X, Y) = p(X)p(Y) \tag{2.14}$$

- **conditionally independent**

$$X \perp Y|Z \iff p(X, Y|Z) = p(X|Z)p(Y|Z)$$

What is $\sum_I P(I, D)$?

- P(I | D)
- P(I)
- P(D | I)
- P(D)

Application for AVs

Problem

TESLA is interested in the deployment of fleet of autonomous vehicles (Avs), they will be commuting mostly on I4. They want to evaluate the risk of driving on this highway. Thus, they want to compute some statistics.

The want to know the times of the day (mornings, afternoons, evenings) where will be more safe for Avs to travel of I4.

Non safe criteria: Highest probability of finding a car driving at more that 100 m/p

The Spring 2020 ML class of Florida Poly got the contract. Now, you have to design the experiment and provide an answer.

Here, a clue

- https://www.youtube.com/watch?v=AgSumauLegs

Problem: Medical Diagnosis

As an example of how to use this rule, consider the following medical diagonsis problem. Suppose you are a woman in your 40s, and you decide to have a medical test for breast cancer called a **mammogram**. If the test is positive, what is the probability you have cancer? That obviously depends on how reliable the test is. Suppose you are told the test has a **sensitivity** of 80%, which means, if you have cancer, the test will be positive with probability 0.8. In other words,

where $x = 1$ is the event the mammogram is positive, and $y = 1$ is the event you have breast cancer. Many people conclude they are therefore 80% likely to have cancer. But this is false! It ignores the prior probability of having breast cancer, which fortunately is quite low:

$$p(y = 1) = 0.004$$

Problem: Medical Diagnosis

Ignoring this prior is called the **base rate fallacy**. We also need to take into account the fact that the test may be a **false positive** or **false alarm**. Unfortunately, such false positives are quite likely (with current screening technology):

$$p(x = 1|y = 0) = 0.1$$

What is is the probability of having cancer if the test was positive ?

- P(y=1 | x =1) = ?

Continuous random variables and Naïve Bayes

- Another world: This is the world of the PDFs

PROBABILITY DENSITY FUNCTIONS

$$P(a < X \leq b) = \int_a^b f(x)dx$$